Evaluating the Robustness of Neural Networks with Adversarial Attacks

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems

Supervisor:
Prof. Dr.-Ing. Matthias Althoff

Advisor:
Lukas Koller, M.Sc.

Research project:
DFG - SPP 2422

Type:
BA

Research area:
Formal Verification of Neural Networks

Programming language:
MATLAB

Required skills:
Machine Learning, Adversarial Attacks

Language:
English, German

Date of submission:
17. October 2024

For more information please contact us:

Phone: +49 (89) 289 - 18140

E-Mail: lukas.koller@tum.de

Website: www.ce.cit.tum.de/air/

Background

Neural networks are great at solving many complex tasks [10, 2, 16, 17]. However, the output of many neural networks is sensitive towards tiny input perturbations [5]. Thus, there is a large field of research centered around training neural networks that are robust against input perturbations [12, 6, 14]. A big challenge is the practical evaluation of the robustness of neural networks. The formal verification of neural networks is computationally hard [8]; thus, even verifying small neural networks is often infeasible. On the contrary, neural networks can be falsified by generating input perturbations that lead to misclassifications, so-called adversarial attacks. Often, adversarial attacks are fast to compute and effective at provoking misclassifications of neural networks [5]. Therefore, adversarial attacks are suitable for evaluating the robustness of neural networks.

Description

There are many methods to compute adversarial attacks [5, 13, 15, 11, 3, 1, 4, 7], which are based on different approaches, e.g., gradient-based [5], optimization-based [3], training neural networks to generate adversarial attacks [1], or reachability analysis [9]. Moreover, there are different threat models and types of attacks: white-box attacks have full knowledge about the neural network, i.e., architecture, parameters, and gradients, while black-box attacks only have restricted knowledge about the neural network; backdoor attacks manipulate the training to embed a backdoor into the behavior of the neural network. Furthermore, several training methods incorporate adversarial attacks to increase their robustness, e.g. [5, 12, 18].

The contributions of this thesis are (i) a comprehensive comparison between different types of adversarial attacks and methods to generate them, (ii) a comparative evaluation of diverse adversarial training strategies, and (iii) a framework to effectively test the robustness of neural networks.

Tasks

1. Literature research on state-of-the-art adversarial attacks.
2. Implementation of selected adversarial attacks.
3. Extensive comparison and evaluation of the implemented attacks.
4. Creation of a framework to effectively test the robustness of neural networks.
5. Training robust neural networks and evaluating their robustness using different adversarial attacks.

References

[1] Shumeet Baluja and Ian Fischer. Learning to attack: Adversarial transformation networks. In *Proc. of the AAAI Conf. on Artificial Intelligence (AAAI)*, 2018.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, 2017.

[4] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *Proc. of the Int. Conf. on Machine Learning (ICML)*, 2020.

[5] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2015.

[6] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Arthur Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 4841–4850, 2019.

[7] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[8] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Int. Conf. on Computer Aided Verification (CAV)*, pages 97–117, 2017.

[9] Niklas Kochdumper, Bastian Schürmann, and Matthias Althoff. Utilizing dependencies to obtain subsets of reachable sets. In *Proc. of the Int. Conf. on Hybrid Systems: Computation and Control (HSCC)*, 2020.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[11] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2017.

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2018.

[13] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Mark Niklas Müller, Franziska Eckert, Marc Fischer, and Martin Vechev. Certified training: Small boxes are all you need. In *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2023.

[15] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE Symposium on Security and Privacy (SP)*, pages 372–387, 2016.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[17] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3960–3969, 2020.

[18] Cihang Xie and Alan Yuille. Intriguing properties of adversarial training at scale. In *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2020.

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems